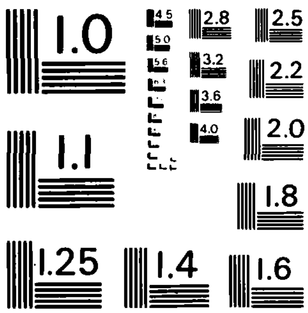
AD-A135 666 A ROBUST MULTIPLE CORRELATION COEFFICIENT FOR THE RANK ANALYSIS OF LINEAR MODELS(U) WESTERN MICHIGAN UNIV KALAMAZOO DEPT OF MATHEMATICS G L SIEVERS SEP 83 TR-69 N00014-78-C-0637 1/1

UNCLASSIFIED F/G 12/1 NL

END
DATE FILMED
1-84
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

6

AD-A135 666

DEPARTMENT OF MATHEMATICS
WESTERN MICHIGAN UNIVERSITY
Kalamazoo, Michigan
49008

A ROBUST MULTIPLE CORRELATION COEFFICIENT FOR THE RANK ANALYSIS OF LINEAR MODELS

by

Gerald L. Sievers

TECHNICAL REPORT NO. 69
September 1983

This research was supported by the Office of Naval Research under contract N00014-78-C-0637 (NR 042-407)

DEPARTMENT OF MATHEMATICS
WESTERN MICHIGAN UNIVERSITY
KALAMAZOO, MICHIGAN 49008

ABSTRACT

A multiple correlation coefficient is discussed to measure the degree of association between a random variable Y and a set of random variables $X_1, \ldots, X_p$. The coefficient is defined in terms of a weighted Kendall's tau, suitably normalized. It is directly compatible with the rank statistic approach of analyzing linear models in a regression, prediction context. The population parameter equals the classical multiple correlation coefficient if the multivariate normal model holds but would be more robust for departures from this model. Some results are given on the consistency of the sample estimate and on a test for independence.

Key Words: Rank statistics, linear models, multiple correlation, robust statistics

1. INTRODUCTION

Consider a context with $p + 1$ random variables Y and $\underline{X} = (X_1, \ldots, X_p)'$. Suppose that Y is viewed as a dependent variable, $X_1, \ldots, X_p$ as independent variables and interest is in measuring the degree of association between Y and $X_1, \ldots, X_p$ as is typical with a multiple correlation parameter.

The classical multiple correlation coefficient $\rho_{Y \cdot X_1 \ldots X_p}$ of the multivariate normal model has many useful properties but it lacks robustness (see Huber (1977)). Its sample estimate is sensitive to outliers and heavier tailed distributions and can be inefficient for nonnormal distributions. An alternate measure is needed which is more robust in such situations.

One important property of $\rho_{Y \cdot X_1 \ldots X_p}$ is that it is the Pearson correlation between Y and a best linear prediction of Y from $\underline{X}$ in the sense of minimum squared error. In this way $\rho_{Y \cdot X_1 \ldots X_p}$ is directly related to regression concepts in interpretation and methodology. This property can be retained in defining a more robust multiple correlation coefficient if the correlation measure and the linear predictor are replaced by more robust choices. This paper will explore such a measure using a linear predictor based on rank estimates of regression coefficients. The measure of association used will be a weighted Kendall's tau parameter which is directly comparable with the rank-regression approach.

Estimates of regression coefficients based on rank statistics have been developed by many authors; in particular, see Jurečková (1971), Jaeckel (1972), McKean and Hettmansperger (1976, 1977) and Sievers (1983) for some of the basic properties and results on their robustness and efficiency. The connection between weighted Kendall's tau statistics and rank regression statistics was mentioned in Sievers (1978).

In a bivariate setting, Kendall's tau is a widely used nonparametric measure of association. Several useful extensions have been discussed for multivariate settings; see Moran (1951), Bobko (1977) and Agresti (1977). This paper will differ by emphasizing the connection to the corresponding regression, prediction problem. A natural population parameter will be used to allow for a direct, meaningful interpretation of sample results. The sample estimate should be highly efficient, in contrast to earlier methods, although a stronger model is needed.

The basic measure of association treated here is a weighted Kendall's tau. The weights will be important in keeping the correlation measure directly compatible with the corresponding regression, prediction concepts and methods. In the regression problem it is known that weights should be used to avoid low efficiency; see Sievers (1978), Scholz (1977). Only in carefully designed experiments where nonrandom, equally spaced values for the independent variables can be set would the weights be

unnecessary, and in such situations multiple correlation issues are usually not important.

2. THE BIVARIATE CASE

This section considers the bivariate case to introduce some ideas and motivate the main definition to follow. Consider a pair of random variables (Y, X) with a nondegenerate bivariate distribution. Let (Y_1, X_1) and (Y_2, X_2) be independent with the same distributions as (Y, X). A widely used nonparametric measure of association is Kendall's tau

$\tau = E(\text{sgn}(X_2-X_1)\ \text{sgn}(Y_2-Y_1))$, where $\text{sgn}(t) = -1, 0, 1$ as $t < 0, = 0, > 0$. The value of τ is in $[-1, 1]$. Following by analogy the Pearson correlation, one could take the absolute value to obtain a multiple correlation coefficient although it is not clear how useful this could be.

The Kendall tau is symmetric in the role of X and Y. However, in the multiple correlation context the variables should be treated asymmetrically, with Y and X playing the part of a dependent and independent variable, respectively. This would relate multiple correlation concepts more directly to regression, prediction concepts as is familiar with the classical $\rho_{Y \cdot X_1 \ldots X_p}$. Moreover, in the regression problem it has been noted in Jaeckel (1972), Scholz (1977) and Sievers (1978) that

the use of weights depending on X is needed to obtain high efficiency in the nonparametric procedure based on Kendall's tau.

These considerations motivate a definition of a correlation coefficient

$$\tau^* = E(|X_2-X_1| \operatorname{sgn}(X_2-X_1) \operatorname{sgn}(Y_2-Y_1))/E(|X_2-X_1|)$$
$$= E((X_2-X_1) \operatorname{sgn}(Y_2-Y_1))/E(|X_2-X_1|),$$

where in the first form, the numerator is a weighted Kendall's tau and the denominator is a suitable norming factor. The use of differences here is natural for parameters based on rank order. It is worth noting that the product-moment correlation coefficient can also be expressed in terms of differences as $\rho = E((X_2-X_1)(Y_2-Y_1))/[E(X_2-X_1)^2 E(Y_2-Y_1)^2]^{1/2}$. Thus τ^* is "in between" ρ and Kendall's tau by replacing one of the variables Y_2-Y_1 by $\operatorname{sgn}(Y_2-Y_1)$.

The parameter τ^* has several desirable properties: $|\tau^*| \leq 1$, τ^* is invariant under linear transformations of the variables, $\tau^* = 0$ if X and Y are independent, $\tau^* = 1$ if Y is a linear function of X with probability one. Also if X and Y have a bivariate normal distribution with correlation ρ then $\tau^* = \rho$. These properties will be discussed in more detail in the multivariate case in the next section.

The definition of τ^* above does not lend itself readily to an extension to higher dimensions and it is not suitable for a multiple correlation since τ^* can be negative. The following change in the definition will allow a natural extension. Let β_* be a value of β minimizing $E(|(Y_2-Y_1) - \beta(X_2-X_1)|)$. Then define

$$\tau = E(\beta_*(X_2-X_1)\ \mathrm{sgn}(Y_2-Y_1))/E(|\beta_*(X_2-X_1)|)$$

if $\beta_* \neq 0$ and $\tau = 0$ if $\beta_* = 0$. Factoring out β_*, it follows that $\tau = \mathrm{sgn}(\beta_*)\tau^*$, so there is at most a sign difference between τ and τ^*. Later it is shown that τ is nonnegative.

3. THE MULTIVARIATE CASE

Consider random variables Y and $\underline{X} = (X_1, \ldots, X_p)'$. Assume they have finite expectations, but otherwise their distribution can be quite arbitrary for some of the material in this section. Of special interest here is the model that specifies the joint cdf of Y and $\underline{X}$ to be of the form

$$F(y - \underline{\beta}_0'\underline{x})H(\underline{x}), \tag{3.1}$$

where F is a univariate cdf, H is a p-dimensional cdf and $\underline{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})'$ is a vector of unknown parameters. In this model the conditional cdf of Y given $\underline{X} = \underline{x}$ is $F(y - \underline{\beta}_0'\underline{x})$. This property appears in the multivariate normal model, but here the F is not assumed normal. No symmetry or centering assumptions are made on F or H. Alternately, this model can be expressed as

$$Y = \underline{\beta}_0'\underline{X} + e, \tag{3.2}$$

where $\underline{X}$ has cdf H, e has cdf F and $\underline{X}$ and e are independent.

Considerations in the bivariate case lead to the following definition of a multiple correlation parameter. Let $(Y_1, \underline{X}_1)$ and $(Y_2, \underline{X}_2)$ be independent, each having the distribution of $(Y, \underline{X})$. Suppose $\underline{\beta}_* = (\beta_{*1}, \dots, \beta_{*p})'$ minimizes

$$E[|(Y_2-Y_1) - \underline{\beta}'(\underline{X}_2-\underline{X}_1)|] = E[|(Y_2 - \underline{\beta}'\underline{X}_2) - (Y_1 - \underline{\beta}'\underline{X}_1)|] \tag{3.3}$$

as a function of $\underline{\beta}$. Then define a multiple correlation parameter by

$$\tau = \frac{E\left[\sum_{k=1}^{p} \beta_{*k}(X_{2k} - X_{1k})\ \mathrm{sgn}(Y_2 - Y_1)\right]}{E\left[\left|\sum_{k=1}^{p} \beta_{*k}(X_{2k} - X_{1k})\right|\right]} = \frac{E\left[\underline{\beta}_*'(\underline{X}_2 - \underline{X}_1)\ \mathrm{sgn}(Y_2 - Y_1)\right]}{E\left[\left|\underline{\beta}_*'(\underline{X}_2 - \underline{X}_1)\right|\right]} \tag{3.4}$$

if $\underline{\beta}_* \neq \underline{0}$ and let $\tau = 0$ if $\underline{\beta}_* = \underline{0}$. In the notation here $\underline{X}_i = (X_{i1}, \dots, X_{ip})'$, $i = 1, 2$. Note that (3.3) is a convex function of $\underline{\beta}$. In most cases of practical interest $\underline{\beta}_*$ will be unique. For ambiguous cases τ will be left undefined.

Note that τ is defined as the weighted Kendalls' tau as modified in Section 2 for Y vs $\underline{\beta}_*'\underline{X}$. The linear function $\underline{\beta}_*'\underline{X}$ can be viewed as a best linear predictor of Y in the sense of minimizing the variation in $Y - \underline{\beta}'\underline{X}$ as measured by the absolute difference of two independent copies. Recall that if z_1 and z_2 are independent copies of a random variable z, then $E(|z_1 - z_2|)$ measures the variation in z (a Gini mean difference parameter). Being of first order, this will be less sensitive to contamination and heavy tails in the distribution in comparison to the square function $E((z_1 - z_2)^2) = 2\ \mathrm{var}(z)$ used in the classical approach. (3.3) is the population analog of the dispersion function used in Sievers (1983).

Remark 3.1. Assume model (3.1). Let G denote the cdf of the difference of two independent random variables each having cdf F and assume G has a unique median. Then $\underline{\beta}_0$ is the unique point minimizing (3.3) and

$$\tau = E[\underline{\beta}_0'(\underline{X}_2 - \underline{X}_1)\ \mathrm{sgn}(Y_2 - Y_1)]/E[|\underline{\beta}_0'(\underline{X}_2 - \underline{X}_1)|].$$

Proof. Under model (3.1) the conditional distribution of $W = Y_2 - Y_1$ given $\underline{X}_2 - \underline{X}_1 = \underline{t}$ has cdf $G(w - \underline{\beta}_0'\underline{t})$. This distribution has a unique median of $\underline{\beta}_0'\underline{t}$ since G has a unique median by assumption and its value is 0 from W being symmetrically distributed about 0. It is well-known that the

median minimizes an expected absolute deviation. Thus for each fixed $\underline{t}$, $E[|W - a||\underline{t}]$ is minimum if $a = \underline{\beta}_0'\underline{t}$ and the result follows. ■

Remark 3.2. If Y and $\underline{X}$ are independent, then $\tau = 0$.

Proof. A conditional argument as in the previous proof shows that $\underline{\beta} = \underline{0}$ minimizes (3.3), although it may not be uni . Regardless, independence implies that the numerator of τ fa tors and the result follows from $E(\mathrm{sgn}(Y_2 - Y_1)) = 0$ by symmetry. ■

The following remark shows an important property; that $\tau = 0$ is equivalent to Y and $\underline{X}$ being independent in model (3.1). The classical parameter $\rho_{Y \cdot X_1 \ldots X_p}$ has this property for the multivariate normal model but not, in general, for nonnormal cases.

Remark 3.3. Assume model (3.1) holds with $\underline{X}$ having a non-degenerate distribution. Then

$$\tau = 0 \Longleftrightarrow \underline{\beta}_0 = 0 \Longleftrightarrow Y \text{ and } \underline{X} \text{ are independent.}$$

Proof. Because of the form assumed for the joint cdf of Y and $\underline{X}$ in model (3.1), Y and $\underline{X}$ are independent if and only if $\underline{\beta}_0 = \underline{0}$. If $\underline{\beta}_0 = \underline{0}$ then $\tau = 0$ by definition. It remains to show that $\underline{\beta}_0 \neq \underline{0}$ implies $\tau \neq 0$.

Assuming $\underline{\beta}_0 \neq \underline{0}$, $T = \underline{\beta}_0'(\underline{X}_2-\underline{X}_1)$ has a nondegenerate distribution with cdf say $L(t)$. Let $W = Y_2-Y_1$. Under model (3.1) the conditional cdf of W given $T = t$ is $G(w-t)$. The numerator of τ is

$$\begin{aligned} E(T\ \mathrm{sgn}(W)) &= E[T\{P(W > 0|T) - P(W < 0|T)\}] \\ &= E[T(1 - 2G(-T))] \\ &= E[T(2G(T) - 1)] \\ &= 2\ E(TG(T)), \end{aligned}$$

using $G(t) + G(-t) = 1$. Then since T is symmetrically distributed about 0, this equals

$$2 \int_0^\infty t[2G(t) - 1]dL(t).$$

The integrand is positive on the range of integration and with T having a nondegenerate distribution the integral is positive. Thus $\tau \neq 0$ as was to be shown. ∎

<u>Remark 3.4</u>. $0 \leq \tau \leq 1$.

Proof: The upper bound follows from

$$\begin{aligned} |E(\underline{\beta}_*'(\underline{X}_2-\underline{X}_1)\ \mathrm{sgn}(Y_2-Y_1))| &\leq E(|\underline{\beta}_*'(\underline{X}_2-\underline{X}_1)\ \mathrm{sgn}(Y_2-Y_1)|) \\ &= E(|\underline{\beta}_*'(\underline{X}_2-\underline{X}_1)|). \end{aligned}$$

For the lower bound, it is enough to show the numerator of τ is nonnegative. Let $W = Y_2 - Y_1$ and $T = \underline{\beta}_*'(\underline{X}_2 - \underline{X}_1)$. Then since $\underline{\beta}_*$ minimizes (3.3), write

$$0 \le E(|W|) - E(|W-T|) = E(|W| - |W-T|)$$

$$= \int_{\substack{w>0\\ t\le w}} t + \int_{\substack{w>0\\ t>w}} (2w-t) + \int_{\substack{w<0\\ t\le w}} (t-2w) + \int_{\substack{w<0\\ t>w}} (-t)$$

$$\le \int_{\substack{w>0\\ t\le w}} t + \int_{\substack{w>0\\ t>w}} t + \int_{\substack{w<0\\ t\le w}} (-t) + \int_{\substack{w<0\\ t>w}} (-t)$$

$$= \int_{w>0} t + \int_{w<0} (-t) = \int t \,\mathrm{sgn}(w),$$

where for simplicity the differential part of the integrals was omitted. This last expression, the numerator of τ, is thus nonnegative. ∎

<u>Remark 3.5</u>. If $Y_2 - Y_1$ and $\underline{\beta}_*'(\underline{X}_2 - \underline{X}_1)$ have the same sign with probability one, then $\tau = +1$.

Proof. Let $W = Y_2 - Y_1$ and $T = \underline{\beta}_*'(\underline{X}_2 - \underline{X}_1)$. The hypothesis implies $\mathrm{sgn}(W) = \mathrm{sgn}(T)$ with probability one. Then the numerator of τ is $E(T\,\mathrm{sgn}(W)) = E(T\,\mathrm{sgn}(T)) = E(|T|)$ which is the denominator of τ. ∎

The following remark shows that for the multivariate normal model, τ is identical to the classical multiple correlation coefficient $\rho_{Y \cdot X_1 \ldots X_p}$. Thus ic would share its many useful properties for this model.

Remark 3.6. If Y and $\underline{X}$ have a multivariate normal distribution, then $\tau = \rho_{Y \cdot X_1 \ldots X_p}$.

Proof. If Y and $\underline{X}$ have a multivariate normal distribution then model (3.1) holds with $\underline{\beta}_0$ being the vector of least-squares regression coefficients. It is well-known that $\rho_{Y \cdot X_1 \ldots X_p}$ is the (Pearson) correlation coefficient of Y and $\underline{\beta}_0'\underline{X}$. This is the same as the Pearson correlation coefficient of the differences $W = Y_2 - Y_1$ and $T = \underline{\beta}_0'\underline{X}_2 - \underline{\beta}_0'\underline{X}_1$. Thus W and T have a bivariate normal distribution with zero means and correlation $\rho_{Y \cdot X_1 \ldots X_p}$. It is straightforward to show that

$$E(T \operatorname{sgn}(W)) = \rho_{Y \cdot X_1 \ldots X_p} \sigma_T \sqrt{2/\pi} \quad \text{and} \quad E(|T|) = \sigma_T \sqrt{2/\pi} ,$$

where σ_T is the standard deviation of T, and the results follows. ∎

Remark 3.7. τ is invariant under nonsingular linear transformations of Y and $\underline{X}$.

Proof. τ depends on Y through a signed difference and it is clear that a linear transformation of Y would have no effect. If $\underline{X}$ is replaced by $\underline{CX}$, where $\underline{C}$ is a $p \times p$ nonsingular matrix, then the $\underline{\beta}_*$ minimizing (3.3) changes to

$(\underline{C}')^{-1}\underline{\beta}_*$. Substituting in (3.4), $((\underline{C}')^{-1}\underline{\beta}_*)'(\underline{C}\underline{X}_i) = \underline{\beta}_*'\underline{X}_i$, $i = 1, 2$, and no change in τ would occur. ∎

4. SAMPLE ESTIMATE OF τ

Let $(Y_1, \underline{X}_1), (Y_2, \underline{X}_2), \ldots, (Y_n, \underline{X}_n)$ be independent replicates of $(Y, \underline{X})$, where $\underline{X}_i = (X_{i1}, \ldots, X_{ip})'$, $1 \leq i \leq n$, and $\underline{X} = (X_1, \ldots, X_p)'$. Define an $n \times 1$ vector $\underline{Y} = (Y_1, \ldots, Y_n)'$, an $n \times p$ matrix $\underline{A} = (X_{ij})$, a parameter vector $\underline{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})'$ and an error vector $\underline{e} = (e_1, \ldots, e_n)'$. If $(Y, \underline{X})$ satisfys model (3.2), then

$$\underline{Y} = \underline{A}\,\underline{\beta}_0 + \underline{e}, \tag{4.1}$$

where the elements of $\underline{e}$ are iid with cdf F, the rows of $\underline{A}$ are iid with cdf H and $\underline{A}$ is independent of $\underline{e}$. An intercept parameter could be added to this model but the procedures here are based on differences and it would cancel out and have no effect.

An estimate of τ can be defined in a natural way as follows. First let $\hat{\underline{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ be a vector that minimizes a dispersion measure of the residuals given by

$$D(\underline{\beta}) = \sum_{i<j} \left|(Y_j - Y_i) - \sum_{k=1}^{p} \beta_k (X_{jk} - X_{ik})\right|$$

$$= \sum_{i<j} \left|(Y_j - \underline{\beta}'\underline{X}_j) - (Y_i - \underline{\beta}'\underline{X}_i)\right|. \tag{4.2}$$

Then define an estimate of τ as

$$\hat{\tau} = \frac{\sum_{i<j} \sum_k \hat{\beta}_k (X_{jk} - X_{ik}) \operatorname{sgn}(Y_j - Y_i)}{\sum_{i<j} \left| \sum_{k=1}^{p} \hat{\beta}_k (X_{jk} - X_{ik}) \right|}$$

$$= \frac{\sum_{i<j} \underline{\hat{\beta}}' (\underline{X}_j - \underline{X}_i) \operatorname{sgn}(Y_j - Y_i)}{\sum_{i<j} \left| \underline{\hat{\beta}}' (\underline{X}_j - \underline{X}_i) \right|} \tag{4.3}$$

if $\underline{\hat{\beta}} \neq \underline{0}$ and $\hat{\tau} = 0$ if $\underline{\hat{\beta}} = \underline{0}$.

The dispersion function (4.2) is a convex, piecewise linear function of $\underline{\beta}$ and as a result there will be a point attaining the minimum, although it may not be unique. This is the same dispersion function used in Sievers (1983) and is algebraically equal to the dispersion function in Jaeckel (1972) and in McKean and Hettmansperger (1976, 1977) when Wilcoxon scores are used. These references point out that the diameter of the set of points attaining the minimum tends to zero asymptotically. Further, $\underline{\hat{\beta}}$ is the rank estimate of the regression scores $\underline{\beta}_0$ and these references contain further results on properties of $\hat{\beta}$, computational methods and more.

The estimate $\hat{\tau}$ has the following properties: $0 \le \hat{\tau} \le 1$, $\hat{\tau} = +1$ if the rank order of the fitted values $\underline{A}\hat{\underline{\beta}}$ is the same as the rank order of $\underline{Y}$ and $\hat{\tau}$ is invariant under nonsingular linear transformations on Y_i and $\underline{X}_i$.

The estimate $\hat{\tau}$ can be expressed in another form to view it more explicitly as a rank statistic. First note the formula

$$\sum_{i<j} (X_{jk} - X_{ik}) \operatorname{sgn}(Y_j - Y_i)$$

$$= \sum_i X_{ik}(2 S_i - (n+1)) = 2 \sum_i (X_{ik} - \overline{X}_k) S_i, \tag{4.4}$$

where S_i is the rank of Y_i among $Y_1, \ldots, Y_n$ and $\overline{X}_k = \sum_i X_{ik}/n$. Using this, the numerator of $\hat{\tau}$ is

$$2 \sum_k \hat{\beta}_k \sum_i (X_{ik} - \overline{X}_k) S_i = 2\, \hat{\underline{\beta}}' \underline{A}_c \underline{S} = 2\, \hat{\underline{Y}}' \underline{S},$$

where $\underline{S} = (S_1, \ldots, S_n)'$, $\hat{\underline{Y}} = \underline{A}_c \hat{\underline{\beta}}$ is the vector of centered fitted values and $\underline{A}_c = (X_{ik} - \overline{X}_k)_{nxp}$ is the centered $\underline{A}$ matrix. (Alternately the rank vector could be centered.) Writing the denominator of $\hat{\tau}$ as $\sum_{i<j} \sum_k \hat{\beta}_k (X_{jk} - X_{ik}) \operatorname{sgn}(\sum_k \hat{\beta}_k (X_{jk} - X_{ik}))$ and applying the same method gives

$$\hat{\tau} = \hat{\underline{Y}}' \underline{S} \,/\, \hat{\underline{Y}}' \hat{\underline{S}}, \tag{4.5}$$

where $\underline{\hat{S}} = (\hat{S}_2, \ldots, \hat{S}_n)$ is the rank vector of $\underline{\hat{Y}}$.

Thus the numerator of $\hat{\tau}$ is $\mathrm{cov}(\underline{\hat{Y}}, \underline{\hat{S}})$ and the denominator is $\mathrm{cov}(\underline{\hat{Y}}, \underline{\hat{S}})$. The covariance of $\underline{\hat{Y}}$ with a permutation of the integers $(1, \ldots, n)$ is maximum when the integers are in the same order as the elements of $\underline{\hat{Y}}$, see Jaeckel (1972). Thus the denominator is the maximum covariance of $\underline{\hat{Y}}$ with a rank vector. This supports the choice of denominator in $\hat{\tau}$, verifys $\hat{\tau} \leq 1$, and shows $\hat{\tau} = +1$ when $\underline{Y}$ and $\underline{\hat{Y}}$ are in the same rank order.

The formula (4.5) suggests an interesting generalization to allow arbitrary scores instead of ranks. Simply replace the rank vectors $\underline{S}$ and $\underline{\hat{S}}$ by the corresponding permutations of a vector of nondecreasing scores $(a_1, \ldots, a_n)$. It appears that such a statistic would have the same properties as $\hat{\tau}$. This will be discussed in a subsequent paper.

5. CONSISTENCY OF $\hat{\tau}$

In this section $\hat{\tau}$ is shown to be a consistent estimate of τ under model (4.1) with some additional regularity conditions:

(C1) The cdf F has an absolutely continuous density function f with $\int (f'/f)^2 f \, dx < \infty$,

(C2) The difference of two independent random variables with cdfs F has cdf G and density function g which is continuous at zero, $g(0) > 0$,

(C3) The random vector $\underline{X}$ has a positive definite variance-covariance matrix $\underline{\Sigma}$,

(C4) There exists a positive δ such that $E[(\underline{X}-\underline{\mu})'(\underline{X}-\underline{\mu})]^{2+\delta} < \infty$, where $\underline{\mu} = E(\underline{X})$.

Some additional notation will be needed for the proofs of this section. Define $T_k(\underline{\beta}) = \sum_{i<j} (X_{jk}-X_{ik}) \operatorname{sgn}[Y_j-Y_i - \underline{\beta}'(\underline{X}_j-\underline{X}_i)]$

and let $\underline{T}(\underline{\beta}) = (T_1(\underline{\beta}), \ldots, T_p(\underline{\beta}))'$. Also let $L(\underline{\beta}) = \sum_{i<j} |\underline{\beta}'(\underline{X}_j-\underline{X}_i)|$. Let $\Delta^* = (1/2\gamma)n^{-3/2}\, \underline{\Sigma}^{-1}\underline{T}(\underline{0})$, where

$\gamma = \int f^2$.

<u>LEMMA 5.1</u>. Assume model (4.1) and conditions C1 - C4. Then if $\underline{\beta}_0 = \underline{0}$,

(i) $n^{-3/2}\underline{T}(\underline{0}) \xrightarrow{\mathcal{D}} N(\underline{0}, (1/3)\underline{\Sigma})$,

(ii) $\hat{\underline{\Delta}} - \underline{\Delta}^* \xrightarrow{P} \underline{0}$, where $\hat{\underline{\Delta}} = \sqrt{n}\, \hat{\underline{\beta}}$, and

(iii) $\hat{\underline{\Delta}} \xrightarrow{\mathcal{D}} N(\underline{0}, (1/12\gamma^2)\underline{\Sigma}^{-1})$.

Note that when $\underline{\beta}_0$ holds, $\sqrt{n}(\hat{\underline{\beta}} - \underline{\beta}_0)$ has the same distribution as $\hat{\Delta}$ when $\underline{\beta}_0 = \underline{0}$ and thus the limiting distribution of (iii).

Proof. The above results were given in Sievers (1983) for the case of nonrandom X_{ij}. The assumptions A1-A8 of that paper will hold almost everywhere in the present context if

$$\max_{1 \le i \le n} |X_{ik} - \bar{X}_{kn}| / \sqrt{n} \longrightarrow 0 \quad \text{a.e.} \quad , \text{ for } \quad 1 \le k \le p, \quad \text{and}$$

$(1/n)\underline{A}_c'\underline{A}_c \longrightarrow \underline{\Sigma}$ a.e. as $n \to \infty$. But these follow from conditions C3 and C4 and Lemma 4.1 of Ghosh and Sen (1971). ■

THEOREM 5.1. Assume model (4.1) and conditions C1-C4. Then $\hat{\tau} \xrightarrow{P} \tau$.

Proof: First consider the case $\underline{\beta}_0 \neq \underline{0}$. Express $\hat{\tau}$ in the form

$$\hat{\tau} = (\hat{\underline{\beta}}' \, \underline{T}(\underline{0})/M)/(L(\hat{\underline{\beta}})/M), \tag{5.1}$$

where $M = \binom{n}{2}$. Similarly write

$$\tau = (\underline{\beta}_0' \, \underline{\mu}^*(\underline{\beta}_0))/\mu(\beta_o),$$

where $\underline{\mu}^*(\underline{\beta}_0)$ is a $p \times 1$ vector with kth element $E[(X_{2k}-X_{1k}) \, \mathrm{sgn}(Y_2-Y_1)]$ and $\mu(\underline{\beta}_0) = E[|\underline{\beta}_0'(\underline{X}_2-\underline{X}_1)|]$.

From Lemma 4.1, it follows that $\hat{\underline{\beta}} \xrightarrow{P} \underline{\beta}_0$. The vector $\underline{T}(\underline{0})/M$ is a vector of U-statistics which converges in probability to $\underline{\mu}^*(\underline{\beta}_0)$ by the usual theory. For the denominator of (5.1), note that $(L(\hat{\underline{\beta}}) - L(\underline{\beta}_0))/M \xrightarrow{P} 0$ since it is bounded above in absolute value by $\max_{1 \le k \le p} |\hat{\beta}_k - \beta_{0k}| \sum_k \sum_{i<j} |X_{jk} - X_{ik}|/M$ and the latter converges to zero in probability. But $L(\underline{\beta}_0)/M$ is a U-statistic converging in probability to $\mu(\underline{\beta}_0)$. It follows that $\hat{\tau} \xrightarrow{P} \tau$ in case $\underline{\beta}_0 \neq \underline{0}$.

Now consider the case $\underline{\beta}_0 = \underline{0}$. The above argument does not apply since both numerator and denominator of $\hat{\tau}$ tend to zero and it is necessary to deal with the rates of convergence. First express (5.1) in terms of $\hat{\underline{\Delta}} = \sqrt{n}\,\hat{\underline{\beta}}$ as

$$\hat{\tau} = (\hat{\underline{\Delta}}'\underline{T}(\underline{0})/M)/(L(\hat{\underline{\Delta}})/M). \qquad (5.2)$$

From Lemma 5.1 (iii), $\hat{\underline{\Delta}}$ is $O_p(1)$ and, as above, $\underline{T}(\underline{0})/M \xrightarrow{P} \underline{\mu}^*(\underline{0}) = \underline{0}$. Thus $\hat{\underline{\tau}} \xrightarrow{P} \underline{0}$ if it is shown that the denominator of (5.2) is bounded away from zero in probability.

To show this let $G_\delta = \{\underline{\Delta} \in R^p\colon ||\underline{\Delta}|| \geq \delta\}$ for $\delta > 0$, where R^p is p-dimensional Euclidean space and $||\cdot||$ the usual distance. Let the boundary be $\mathcal{B}_\delta = \{\underline{\Delta} \in R^p\colon ||\underline{\Delta}|| = \delta\}$. By Lemma 5.1 (iii), $P(\hat{\underline{\Delta}} \in G_\delta)$ can be made arbitrarily close to one for all n sufficiently large by taking δ sufficiently small. Now for any fixed $\underline{X}_1, \ldots, \underline{X}_n$, $L(\underline{\Delta})$ is nonnegative, convex, $L(\underline{0}) = 0$ and so for any $\underline{\Delta}' \in G_\delta$ there exists $\underline{\Delta} \in \mathcal{B}_\delta$ such that $L(\underline{\Delta}) \leq L(\underline{\Delta}')$. Thus if $\hat{\underline{\Delta}} \in G_\delta$, $L(\hat{\underline{\Delta}})/M \geq \inf_{\underline{\Delta} \in \mathcal{B}_\delta} L(\underline{\Delta})/M$ and it will be sufficient to show the latter is bounded away from zero in probability.

To accomplish this a compactification argument can be used. $\mathcal{B}_\delta$ is a compact set. For any $\underline{\Delta}, \underline{\Delta}' \in \mathcal{B}_\delta$, it can be shown that $|L(\underline{\Delta}) - L(\underline{\Delta}')|/M \leq ||\underline{\Delta} - \underline{\Delta}'||\,V$, where $V = \Sigma_k\ \Sigma_{i<j}|X_{jk}-X_{ik}|/M$ converges in probability. Also $L(\underline{\Delta})/M \xrightarrow{P} \mu(\underline{\Delta})$ for any fixed

point $\underline{\Delta}$ and therefore uniformly for any finite set of points $\underline{\Delta}$. Finally, use the fact that $\inf_{\underline{\Delta} \in \mathcal{B}_\delta} \mu(\underline{\Delta}) > 0$, since $\mu(\underline{\Delta})$ is nonnegative, convex, $\mu(\underline{0}) = 0$ and $\mu(\underline{\Delta}) = 0$ for some $\underline{\Delta} \neq \underline{0}$ would contradict the assumption of a positive definite $\underline{\Sigma}$ matrix. ∎

6. A TEST OF INDEPENDENCE

In this section a test of the hypothesis of independence is considered for model (4.1). In view of Remark 3.3, this is the hypothesis H_0: $\tau = 0$ (or $\underline{\beta}_0 = \underline{0}$). The test will be based on the numerator of $\hat{\tau}$, viewing its denominator as basically a norming factor. The distribution theory for the numerator of $\hat{\tau}$ is readily available from the results of Section 5.

From (4.3) and (4.4), the numerator of $\hat{\tau}$ is $\underline{\hat{\beta}}'\underline{T}(\underline{0}) = 2\,\underline{\hat{Y}}'\underline{S}$, where $\underline{\hat{Y}} = \underline{A}_c\underline{\hat{\beta}}$ is the centered vector of fitted values and $\underline{S}$ is the rank vector of $\underline{Y}$. The proposed test of the hypothesis H_0: $\tau = 0$ vs H_1: $\tau > 0$ is to reject H_0 if $Q > \chi^2_{\alpha,p}$, where $Q = (12\hat{\gamma}/n)\,\underline{\hat{Y}}'\underline{S}$, $\hat{\gamma}$ is a consistent estimate of $\gamma = \int f^2$ (see McKean and Hettmansperger (1976, 1977), Sievers and McKean (1983)) and $\chi^2_{\alpha,p}$ is the quantile of order $1 - \alpha$ of a chi-square distribution with p degrees of freedom.

<u>THEOREM 6.1</u>. Assume model (4.1) and conditions C1-C4. Then under H_0, Q has a limiting chi-square distribution with p degrees of freedom.

Proof. It is sufficient to replace $\hat{\gamma}$ by γ and consider, with notation from Section 5,

$$(12\gamma/n)\ \hat{\underline{Y}}'\underline{S} = 6\ \gamma\hat{\underline{\Delta}}'[n^{-3/2}\underline{T}(\underline{0})] = 12\gamma^2\ \hat{\underline{\Delta}}'\Sigma\underline{\Delta}^*. \qquad (6.1)$$

Using Lemma 5.1, this has the same limiting distribution as $12\gamma^2\ \hat{\underline{\Delta}}'\Sigma\hat{\underline{\Delta}}$, which is $\chi^2(p)$. ∎

McKean and Hettmansperger (1976, 1977) have proposed a test of the equivalent hypothesis H_0: $\underline{\beta}_0 = \underline{0}$ based on a drop in dispersion for the case of fixed X_{ij}. In the notation here, this statistic is $(12\hat{\gamma}/n)(D(\underline{0}) - D(\hat{\underline{\beta}}))$, where D is given in (4.2). The asymptotics of Section 5 can be used to show this statistic is asymptotically equivalent to Q and in this sense there is agreement between the tests of $\tau = 0$ and $\underline{\beta}_0 = \underline{0}$. Another test statistic, asymptotically equivalent to Q, arises by replacing $\hat{\underline{\Delta}}$ by $\underline{\Delta}^*$ in (6.1), namely $3n^{-3}\underline{T}(\underline{0})'\underline{\Sigma}^{-1}\underline{T}(\underline{0})$. This statistic has the advantage of not requiring an estimate of the scale parameter γ.

REFERENCES

Agresti, A. (1977), "A Coefficient of Multiple Association Based on Ranks," Communications in Statistics - Theory and Methods-A, 6, 1341-1359.

Bobko, P. (1977), "A Note on Moran's Measure of Multiple Rank Correlation," Psychometrika, 42-311-314.

Ghosh, M., and Sen, P.K. (1971), "On a Class of Rank Order Tests For Regression With Partially Informed Stochastic Predictors," The Annals of Mathematical Statistics, 42, 650-661.

Hettmansperger, T.P. and McKean, J.W. (1977), "A Robust Alternative Based on Ranks to Least Squares in Analyzing Linear Models," Technometrics, 19, 275-284.

Huber, P.J. (1977), "Robust Covariances," In Statistical Decision Theory and Related Topics II, ed. S.S. Gupta and D.S. Moore, New York: Academic Press, 165-191.

Jaeckel, L.A. (1972), "Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals," Annals of Mathematical Statistics, 43, 1449-1458.

McKean, J.W. and Hettmansperger, T.P. (1976), "Tests of Hypotheses Based on Ranks in the General Linear Model," Communications in Statistics - Theory and Methods - A, 8, 693-709.

Moran, P.A.P. (1951), "Partial and Multiple Rank Correlation," Biometrika, 38, 26-32.

Scholz, F.W. (1977), "Weighted Median Regression Estimates," Institute of Mathematical Statistics Bulletin, 6, 44.

Sievers, G.L. (1983), "A Weighted Dispersion Function For Estimation In Linear Models," Communications in Statistics - Theory and Methods - A, 12, 1161-1179.

Sievers, G.L., and McKean, J. (1983), "On the Robust Rank Analysis of Linear Models With Nonsymmetric Error Distributions," unpublished manuscript.

Sievers, G.L. (1978), "Weighted Rank Statistics for Simple Linear Regression," Journal of the American Statistical Association, 73, 628-631.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER 69	2. GOVT ACCESSION NO. AD-A135 666	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Robust Multiple Correlation Coefficient for the Rank Analysis of Linear Models		5. TYPE OF REPORT & PERIOD COVERED Technical Report 1983
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Gerald L. Sievers		8. CONTRACT OR GRANT NUMBER(s) N00014-78-C-0637
9. PERFORMING ORGANIZATION NAME AND ADDRESS Western Michigan University Kalamazoo, Michigan 49008		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-407
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics and Probability Program		12. REPORT DATE September 1983
		13. NUMBER OF PAGES 23
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Please see next page

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Please see next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract

ABSTRACT

A multiple correlation coefficient is discussed to measure the degree of association between a random variable Y and a set of random variables $X_1, \ldots, X_p$. The coefficient is defined in terms of a weighted Kendall's tau, suitably normalized. It is directly compatible with the rank statistic approach of analyzing linear models in a regression, prediction context. The population parameter equals the classical multiple correlation coefficient if the multivariate normal model holds but would be more robust for departures from this model. Some results are given on the consistency of the sample estimate and on a test for independence.

19. Key Words: Rank statistics, linear models, multiple correlation, robust statistics

FIL